A Machine Learning based Approach for Determining Consumer Purchase Intention using Tweets

Sumith Pevekar
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, India

Naresh Alwala
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, India.

Aayush Pandey
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, India.

Prof. Prakash Parmar
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, India.

Abstract—**The ecommerce business, and more especially customers buying things online, has seen a considerable increase recently. People are increasingly posting online about whether they want to buy the product or are inquiring whether they should buy it or not. A lot of research has gone into determining a user's buying tendencies and, more crucially, the elements that affect whether the user will buy the product or not. Twitter is one such network that has grown in popularity in recent years. We will investigate the problem of recognizing and forecasting a user's purchase intention for a product in this study. After applying various text analytical models to tweets data, we discovered that it is possible to predict whether a user has expressed a purchase intention for a product or not, and after further analysis, we discovered that people who initially expressed a purchase intention for a product have in most cases also purchased the product.**

Keywords—*Natural Language Processing, Product, Purchase Intention, Tweets, Twitter*

I. INTRODUCTION

Social networking is a real time micro-blogging like twitter and Facebook where users always receive, send messages, and share information [1]. As far as conversations about consumer products is concerned it is more prominent on Twitter microblogging application. Tweets can be fetched from twitter regarding shopping websites, or any other twitter pages like some business, mobile brands, cloth brands, live events like sport match, election etc. get the polarity of it. These results will help the service provider to find out about the customers view toward their products [2]. Twitter has become a popular online social networking site for sharing real-time information on recent and popular events. At present, lot of research is being conducted for efficiently utilizing the large amount of information posted on Twitter by different users. The research includes areas like detecting communities in social networks, methods for summarization, analyzing tweets and so on [3]. Social media platform like twitter where users can post their tweets in 280 characters. Because of the limited number of characters in tweets, it becomes easy for the sentiment analysis. On Twitter 550 million of tweets are posted daily. Twitter also represents all age group people and a fair representation of gender. Therefore, the sentiment analysis of twitter data becomes somewhat general sentiments of society [4]. In this paper, we use different machine learning algorithms to analyze the purchase intention of the user. We have used algorithms like Support Vector Machine and LSTM. We will compare the accuracy and precision of all these methods to find out which one works the best.

II. LITERATURE REVIEW

Several research studies have been conducted to examine the buying habits of online consumers. Only a few, however, have addressed the customers' intent to purchase things. Ramanand et al. (Ramanand, Bhavsar, and Pedanekar 2010) consider the challenge of finding 'buy' wishes from product reviews in their study on identifying wishes from texts [5]. These wishes can range from product suggestions to a desire to purchase a product. To detect these two types of wishes, they employed language principles. While rule-based approaches to recognizing wishes are successful, their reach is insufficient, and they are difficult to broaden.
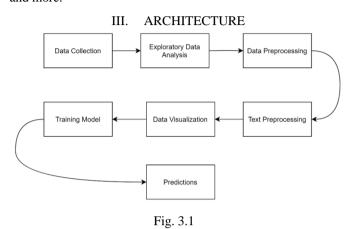
Finding wishes in product reviews is comparable to recognizing purchasing intentions. Instead of using a rule-based approach, we offer a machine learning model that uses generic data acquired from tweets.

Previous research has shown that tweets can be processed using Natural Language Processing (NLP) and Named Entity Recognition (NER) (Li et al., 2012). (Liu et al., 2011) [6]. However, applying NER to tweets is problematic because to the widespread usage of acronyms, misspelt words, and grammatical problems in tweets. Finin et al. (2010) used crowdsourcing to annotate identified items in tweets. These techniques were also used in other studies to apply sentiment analysis to tweets [7].

The sentiment 140 API (a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets), the TweetNLP library (a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets), unigrams, bigrams, and stemming are some common

preprocessing techniques commonly used for twitter data. There are also dictionary-based approaches, such as using the textBlob library (TextBlob is a Python (2 and 3) library for textual data processing.) It provides a unified API for common natural language processing (NLP) operations like part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.

III. ARCHITECTURE

Fig. 3.1

IV. METHODOLOGY

A. Data Collection

The tweets containing products name were gathered from the twitter using Twitter API to create the dataset. There was no publicly available dataset which met our purpose. So, we collected data that showed the purchase intention for products. (Figure 2).

B. Exploratory Data Analysis

Exploratory data analysis, often known as E.D.A., is a key phase in exploring and investigating diverse data sets and summarizing their significant properties, sometimes employing various data visualization approaches. It makes it easier for a data analyst to see trends, spot abnormalities, test theories, and draw conclusions about the best approach to monitor data sources to generate more precise answers.

C. Data Preparation

Data pre-processing is a technique for converting raw data into a usable and efficient format. Training a model with a dataset which is not pre-processed can lead to poor results, therefore Data pre-processing is an important step in machine learning algorithms. In this we have removed unwanted columns to reduce dimensionality. Further we have removed the tuples having undefined class label and normalized the class labels.

D. Text Preprocessing

Simply said, pre-processing your text implies converting it into a format that is predictable and analyzable for your task. General tweets contain various symbols like '@', etc. Such symbols are removed before tokenizing and all bad symbols are removed. Stop words are also removed from the dataset. Next, we have tokenized the text since neural networks work by performing computation on numbers, passing in a bunch of words won't work. This data can now be used for   training the model as well as visualization.

E. Text Visualization

Text visualization is the process of extracting information from the raw text and perform different analysis to derive important structure and pattern out of the information. This is where modern natural language processing (NLP) tools come in. They can capture prevailing moods about a particular topic or product (sentiment analysis), identify key topics from texts (summarization/classification), or amazingly even answer context-dependent questions (like Siri or Google Assistant). Their development has provided access to consistent, powerful, and scalable text analysis tools for individuals and organizations.

For this project we used seaborn library along with Word Cloud to visualize the sentiment polarity as well as subjectivity of tweets in dataset. We used Text Blob library to measure the sentiment and subjectivity count of each tweet.
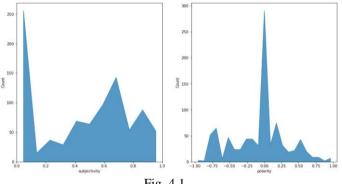

Fig. 4.1

We also used the word cloud library to identify the most recurring words for purchase intended tweet as well as non-intended tweets.

Positive Tweets:

Fig. 4.2

Negative Tweets:

Fig. 4.3

F. Data Training

Machine Learning algorithms learn from data. They find relationships, develop understanding, make decisions, and evaluate their confidence from the training data they're given. And the better the training data is, the better the model performs.

We used the Scikit Learn library to annotate the class labels into binary values using label encoder. We also normalized the data using Minmax scaler to evenly distribute the datapoint values around mean value. Also, vectorized the text into array of sequences using NLTK library.

Finally, after normalization we split the dataset into training and testing in the ratio of 4:1 respectively and used the Keras Long Short-Term Memory (LSTM) model which is a recurring neural network. For LSTM model we added 3 layers using softmax activation at output layer. We tuned the LSTM model into 7 epochs or passes with a batch size of 64.

V. RESULT

We implemented 2 machine learning models namely support vector machine (SVM) and deep learning model long short memory network (LSTM) and achieved an average accuracy of 76.9% which is more accurate than the built-in sentiment analysis model of NLTK library.
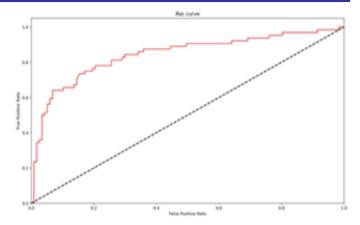

Fig. 5.1 SVM Result

Fig. 5.2 ROC Curve

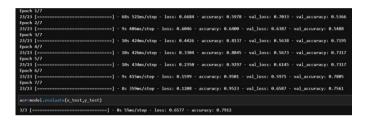Fig. 5.3 LSTM Accuracy

VI. CONCLUSION AND FUTURE SCOPE

Till date we have successfully created our own database for the model using the web crawler software to extract the tweets from twitter and annotated the data by classifying the tweets as a purchase intent tweet or not. We had to create our own dataset because there does not exist a publicly available dataset for purchase intention based on twitter tweets.

Also, we have implemented 2 deep learning models with an average accuracy of 76.9%. The future of trained sentiment analysis will delve deeper beyond the concept of classifying the customer into different segments. As a result, these models are becoming necessary for the businesses to survive in a competitive market

VII. REFERENCES

[1] Rehab S. Ghaly, Emad Elabd, Mostafa Abdelazim Mostafa, Tweet's classification, hashtags suggestion and tweets linking in social semantic web, IEEE, SAI Computing Conference (SAI), 2016, pp. 1140-1146.

[2] Shital Anil Phand, Jeevan Anil Phand , Twitter sentiment classification using Stanford NLP, 1st International Conference on Intelligent Systems and Information Management (ICISIM), IEEE, 2017, pp. 1-5.

[3] Swati Powar, Subhash Shinde, Named entity recognition and tweet sentiment derived from tweet segmentation using Hadoop, 1st International Conference on Intelligent Systems and Information Management (ICISIM), IEEE, 2017, pp. 194 - 198.

[4] Lokesh Mandloi, Ruchi Patel, Twitter Sentiments Analysis Using Machine Learning Methods, International Conference for Emerging Technology (INCET), IEEE, 2020, pp. 1-5.

[5] J. Ramanand, Krishna Bhavsar, Niranjan Pedanekar, Wishful thinking: finding suggestions and 'buy' wishes from product reviews, Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, pp. 54-61.

[6]   Nanyun Peng, Mark Dredze, Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings, Conference: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 548 – 554

[7]   Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, Mark Dredze, Annotating Named Entities in Twitter Data with Crowdsourcing, Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, ACM, 2010,   pp. 80-88